Project 2 (15%)
STQD6014
SEMESTER 1 2023/2024

As a skilled Data Scientist, your role involves uncovering and communicating the narratives hidden within datasets to stakeholders through visualizations. To initiate this process, please visit kaggle.com and select any **RAW DATA SET** of your preference, downloadable in **.CSV format** for the initial data exploration phase. Your tasks encompass:

1. Data Cleaning:
- Thoroughly clean the **RAW DATA** to ensure accuracy and reliability in subsequent analyses.

2. Visualizations:
- Generate a **minimum of five distinct visualization plots**, employing techniques such as pie charts, bar plots, box plots, etc.
- Utilize **Matplotlib**, **Seaborn** or other relevant packages of your choice for creating these visualizations.

3. Insights and Explanations:
- Accompany each plot with a **compelling insight and explanation** that is both relevant and significant. Provide clarity on the story the data tells. You are encouraged to substantiate your observations with other published information online.

4. **Google Colab Notebook** or **Jupyter Notebook**:
- Use the **Google Colab Notebook** or **Jupyter Notebook** as the platform for your assignment.
- Structure the Notebook like a book, encompassing sections such as:
  - **Introduction**: Present the purpose and context of the analysis.
  - **Problem Statement**: Clearly define the problem or question you aim to address.
  - **Results and Discussion**: Showcase your visualizations, insights, and explanations.
  - **Conclusion**: Summarize key findings and any implications.

5. Tools:
- You have the flexibility to choose relevant packages and tools for analysis, with a suggestion to consider **Seaborn** for visualization.

6. Deadline:
- The submission deadline for the Notebook is set for **2024-01-16**.

7. Submission:
- Share your completed **Google Colab Notebook** or **Jupyter Notebook** with me at the email address bernardlkb@ukm.edu.my.

Approach this assignment with the mindset of producing a comprehensive and well-organized document. Your goal is to make the data and its story accessible and understandable to stakeholders.

Criteria	**Marks**		
Reproducibility	**3** The notebook is 100% reproducible	**2** The notebook is reproducible with a few missing steps	**1** The notebook is not reproducible
Plots	**10** All the plots are i. suitable, ii. easy to understand iii. observations are properly explained	**7** Some of the plots are i. suitable, ii. easy to understand iii. observations are properly explained	**5** The plots are i. not suitable, ii. hard to understand iii. observations are poorly explained
Notebook presentation	**2** The overall notebook is i. properly structured, ii.each section neatly organized, iii. easy to follow	**1** Part of the notebook is i. properly structured, ii.each section neatly organized, iii. easy to follow	**0** The notebook is i. poorly structured, ii. each section is not organized, iii. hard to follow